RADC-TR-86-135
In-House Report
August 1986

COMPARISON OF THE EFFECTS OF BROAD-BAND NOISE ON SPEECH INTELLIGIBILITY AND VOICE QUALITY RATINGS

Caldwell P. Smith

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, NY 13441-5700

This report has been reviewed by the RADC Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-86-135 has been reviewed and is approved for publication.

APPROVED:

J.P. VETRANO
Chief, COMSEC Engineering Office
Electromagnetic Sciences Division

APPROVED:

ALLAN C. SCHELL
Chief, Electromagnetic Sciences Division

FOR THE COMMANDER:

JOHN A. RITZ
Plans and Programs Division

DESTRUCTION NOTICE - For classified documents, follow the procedures in DOD 5200.22-M, Industrial Security Manual, Section II-19 or DOD 5200.1-R, Information Security Program Regulation, Chapter IX. For unclassified, limited documents, destroy by any method that will prevent disclosure of contents or reconstruction of the document.

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (EEV) Hanscom AFB MA 01731-5000. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document requires that it be returned.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) RADC-TR-86-135	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Rome Air Development Center 6b. OFFICE SYMBOL (If applicable) EEV	7a. NAME OF MONITORING ORGANIZATION Rome Air Development Center (EEV)
6c. ADDRESS (City, State, and ZIP Code) Hanscom AFB Massachusetts 01731-5000	7b. ADDRESS (City, State, and ZIP Code) Hanscom AFB Massachusetts 01731
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Rome Air Development Center 8b. OFFICE SYMBOL (If applicable) EEV	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) Hanscom AFB Massachusetts 01731-5000	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 33401F; PROJECT NO. 7820; TASK NO. 03; WORK UNIT ACCESSION NO. 01
11. TITLE (Include Security Classification) Comparison of the Effects of Broad-Band Noise on Speech Intelligibility and Voice Quality Ratings	
12. PERSONAL AUTHOR(S) Caldwell P. Smith	
13a. TYPE OF REPORT In-House 13b. TIME COVERED FROM 85-8-1 TO 86-3-31	14. DATE OF REPORT (Year, Month, Day) 1986 August 15. PAGE COUNT 40
16. SUPPLEMENTARY NOTATION	

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Voice communication; Speech perception; Intelligibility; Articulation index
17	02		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Results of a study of effects of broad-band noise on speech intelligibility and voice quality are presented, with a comparison of three methods for evaluating speech signals: the Diagnostic Rhyme Test (DRT) for speech intelligibility, the Diagnostic Acceptability Measure (DAM) test for voice quality and acceptability, and the Mean Opinion Score (MOS), also for evaluating speech quality and acceptability.

Speech samples were combined with broad-band noise, with accurate calibration of speech-to-noise energy ratios by using a new measurement algorithm developed under RADC sponsorship. Four scramblings (randomizations) of the Diagnostic Rhyme Test were prepared with each of three male speakers, for assessing speech intelligibility. Connected speech samples from the three speakers were prepared for assessing voice quality and acceptability. The processed speech samples were digitally recorded for subsequent presentation to listener crews.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Caldwell P. Smith	22b. TELEPHONE (Include Area Code) (617) 377-3281 22c. OFFICE SYMBOL RADC/EEV

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

19. (Contd)

The resulting intelligibility scores were used in constructing the relationship between Diagnostic Rhyme Test scores and values of the articulation index. From this relationship it was possible to estimate from previous standardized intelligibility data, the percent accuracy with which listeners would be expected to receive stereotyped operational messages ("sentences known to listeners") corresponding to various DRT intelligibility scores. For example, it was estimated that a DRT intelligibility score of 70, which has been estimated to be the threshold value below which intelligibility would be unacceptable, corresponds approximately to the value of the articulation index for which listeners would be expected to get 92 percent correct "sentences known to listeners."

Intelligibility scores and voice quality ratings were also used in constructing regression models relating scores with speech-to-noise energy ratios. Category scales for intelligibility scores, and voice quality ratings, were compared on a basis of their common relation to the scale of S/N ratios. It was determined that categories (for example, "very good", "fair", and so on) for intelligibility scores and for DAM voice quality ratings were in rough agreement at the top and bottom of the range studied here (30 dB S/N ratio to 6 dB S/N ratio) but agreed poorly in the middle of the range. Categories for mean opinion scores were in poor agreement with the other two category scales.

Preface

The method for calibrating speech-to-noise energy ratios and the software for using that algorithm were devised by J. T. Sims. The voice recordings of speech with noise were prepared by Paul Gatewood. The assistance of Marianne Bahia, John Fisher, John Olender, Fred Reinhard, Denis Robitaille and Luigi Spagnuolo in making this study possible is gratefully acknowledged.

Contents

Illustrations

Tables

Comparison of the Effects of Broad-Band Noise on Speech Intelligibility and Voice Quality Ratings

I. INTRODUCTION

In the past few years, increased attention has been focused on effects of acoustic background noise in degrading performance of digital voice communications processors. In order to expand the information on this problem, a number of noise environments of special interest including jet and prop aircraft cabin noise, typical office noise, noise in shipboard environments, and the background noise in certain vehicles were measured and recorded, and subsequently simulated in sound rooms for the purpose of preparing standardized speech recordings representative of the effects of those noise environments, for assessing speech intelligibility and voice quality of various digital voice communications processors.

Those studies did not attempt to address the effects of noise environments on listeners, for several reasons. Designers of digital voice processors and pre-processors have virtually no options available for modifying their algorithms to remedy that problem. In many cases, appropriate headphones and ear protectors provide adequate solutions to the problem of noisy listener environments. Of particular importance, speech testing to assess voice quality and naturalness is most critically conducted when listeners are in a quiet environment, since when listeners are placed in a noisy acoustic environment for the purpose of conducting

(Received for publication 7 August 1986)

speech tests, the noise can tend to mask distortions and make systems having degraded voice quality sound more acceptable.

Using recorded speech test materials representing the various noise environments that were simulated, it has been possible to conduct intelligibility and voice quality tests of voice processors with a high degree of reliability and repeatibility of test results. However, it was found that there were wide variations, as much as 10 to 12 dB, in the speech-to-noise energy ratios of different talkers used in these simulations, even under conditions of close control of sound pressure levels, careful phasing of transducers to obtain uniform noise fields, and close attention to details of microphone placement, instructions to talkers, and so on.

Facing a need to accurately calibrate dozens of recordings of speech by multiple talkers in various noise environments, a study was funded under which a contractor worked in the Rome Air Development Center speech test and evaluation facility to develop a measurement algorithm that might be used to facilitate accurate, efficient measurement and calibration of speech-to-noise energy ratios. The successful result of that study has now been published in the literature,[1] and the measurement algorithm is now being used to accurately calibrate each speaker's speech recordings in the speech test and evaluation library of recordings used by the Department of Defense Digital Voice Processor Consortium. The algorithm was also used to calibrate speech-to-noise energy ratios in this pilot study.

2. SPEECH MATERIALS FOR ASSESSING EFFECTS OF BROAD-BAND NOISE

This study utilized existing recordings of three male speakers in a quiet non-reverberant acoustic environment, using a high-quality (Altec 659A) dynamic microphone in a close-talking position approximately 6 cm from the lips. The reproduced speech signals were electrically mixed with white noise from a broad-band noise generator and both were low-pass filtered at 4 kHz. Speech levels were standardized using the measurement algorithm of Brady,[2] and using the calibration method of Sims, the speech-to-noise energy ratios were successively set at 6 dB, 12 dB, 18 dB, 24 dB, and 30 dB. At each S/N ratio high quality digital audio recordings were prepared with a Sony PCM F1 digitizer, using sentence lists for the purpose of assessing voice quality and acceptability and four scramblings

1. Sims, J.T. (1985) A speech-to-noise ratio measurement algorithm, J. Acoust. Soc. Am., 78(No. 5):1671-1674.

2. Brady, P.T. (1968) Equivalent peak level: a threshold-independent speech-level measure, J. Acoust. Soc. Am., 44:695-699.

(randomizations) of the Diagnostic Rhyme Test (DRT) for assessing speech intelligibility, for each of the three male speakers.

3. LISTENER TESTS

Evaluations of the speech test materials were performed "blind" by a contractor, Dynastat Inc., in which ten-member listener crews were presented the reproduced digital recordings at an optimum listening level over headphones in a sound room. Listener judgements of voice quality and acceptability were assessed independently by two methods: the Diagnostic Acceptability Measure (DAM) test of Voiers,[3] and by mean opinion scores (MOS).[4]

4. INTELLIGIBILITY TESTS RESULTS

Results of these intelligibility tests are summarized in Figure 1. Using three talkers and four replications of the test at each of the five S/N ratios provided 12 speaker scores at each S/N ratio, or 60 scores in all. The scatter diagram of scores shows the variation in scores that occurred at each S/N ratio, and an exaggeration of a typical tendency for dispersion of scores to vary inversely with the average score.

Several regression models were calculated for the relationship between intelligibility scores and S/N ratio, which led to a choice of the equation and regression line shown plotted in Figure 1, expressing intelligibility in relation to the reciprocal of the S/N ratio in dB. That particular regression model resulted in a value of $r^2 = 0.91$ and a standard error of estimate of 2.48. The regression equation is obviously not useful for extrapolating outside the range from 6 to 30 dB S/N (on this scale, 0 dB S/N is at infinity). However, the regression model is suited for the purpose intended here, of relating this data to a scale of categories of speech intelligibility scores.

---

3. Voiers, W.D. (1977) Diagnostic acceptability measure for speech communications systems, IEEE Proc. ICASSP 77CH1197-3 ASSP, pp. 204-207.

4. CCITT (1981) Telephone Transmission Quality: Recommendations of the P Series, Yellow Book Vol. V, ITU, Geneva.

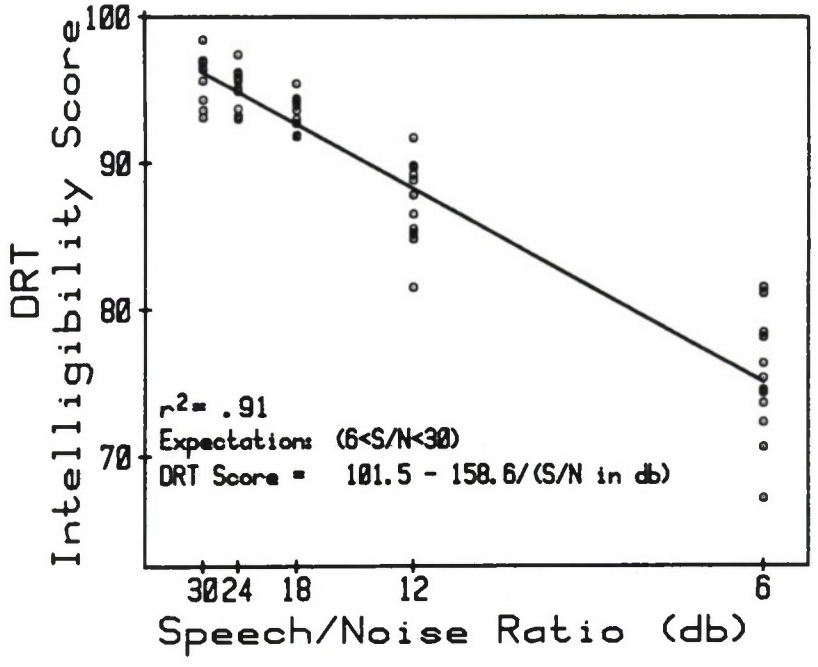

Figure 1. Scatter Diagram of Overall Intelligibility Scores by Speakers, for Speech in Broad-band Noise, With a Regression Model Based on the Reciprocal of S/N Ratio

The category scale for DRT intelligibility scores was established to assist in the interpretation of intelligibility scores by users and planners of digital voice communications systems. Intelligibility scores are classified in terms of eight categories, ranging from "excellent" to "unacceptable", based on the ranges illustrated in Figure 2. This category scale, which has been published previously,[5,6] rates intelligibility scores below 70 as "unacceptable". However, there has been some evidence that highly stereotyped messages well-known to talkers and listeners can be successfully exchanged over a telephony channel even under conditions such that the average intelligibility of the channel (assessed with the Diagnostic Rhyme Test) is below 70.

---

5. Smith, C. P. (1983) Narrowband (LPC-10) Vocoder Performance Under Combined Effects of Random Bit Errors and Jet Aircraft Cabin Noise, RADC-TR-83-293, AD A141333, Rome Air Development Center, Griffiss AFB, N. Y.

6. Smith, C. P. (1983) Relating the performance of speech processor to the bit error rate, Speech Technology 2(No. 1):41-53.

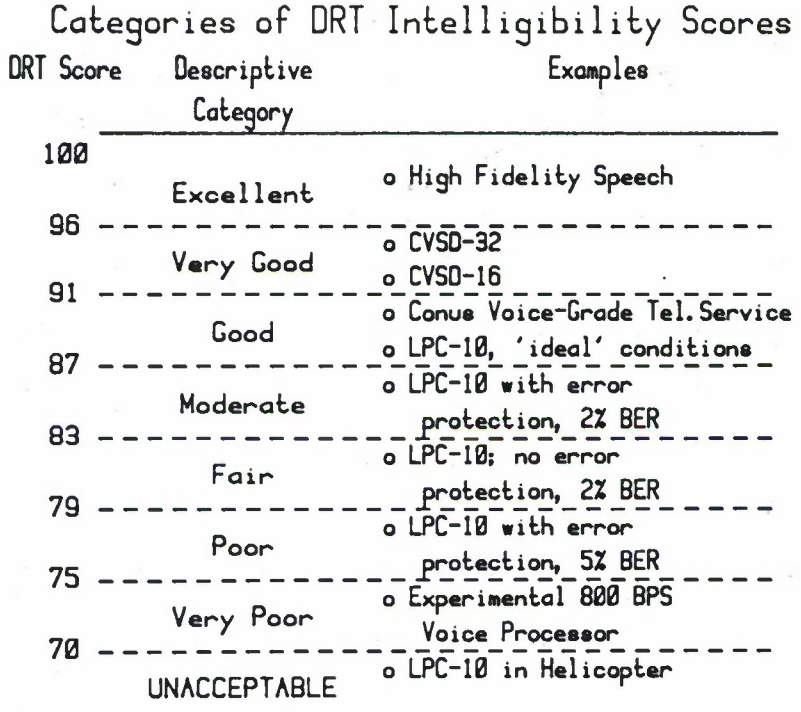

Figure 2. Category Scale for Diagnostic Rhyme Test Intelligibility Scores, With Examples of Voice Processor Categories

When the category scale is combined with the intelligibility scores and regression model obtained in this study, the result presented in Figure 3 is obtained. A 30 dB S/N ratio resulted in scores distributed about equally in the "excellent" and the "very good" categories, with the average value approximately at the boundary between these categories.

The average score at 24 dB S/N ratio was in the "very good" category, with a few speaker scores in the "excellent" category.

All of the scores clustered in the "very good" category for the 18 dB S/N condition.

Dispersion of the individual scores was noticeably increased at 12 dB S/N ratio, with individual scores ranging from "very good" to "fair", with the average value falling in the "good" category. Still greater dispersion was evident at 6 dB S/N, individual scores ranging from "fair" (two), to "poor" (four), to "very poor" (five) to "unacceptable" (one). The average intelligibility obtained at this S/N ratio was approximately at the boundary between the "poor" and "very poor" categories.

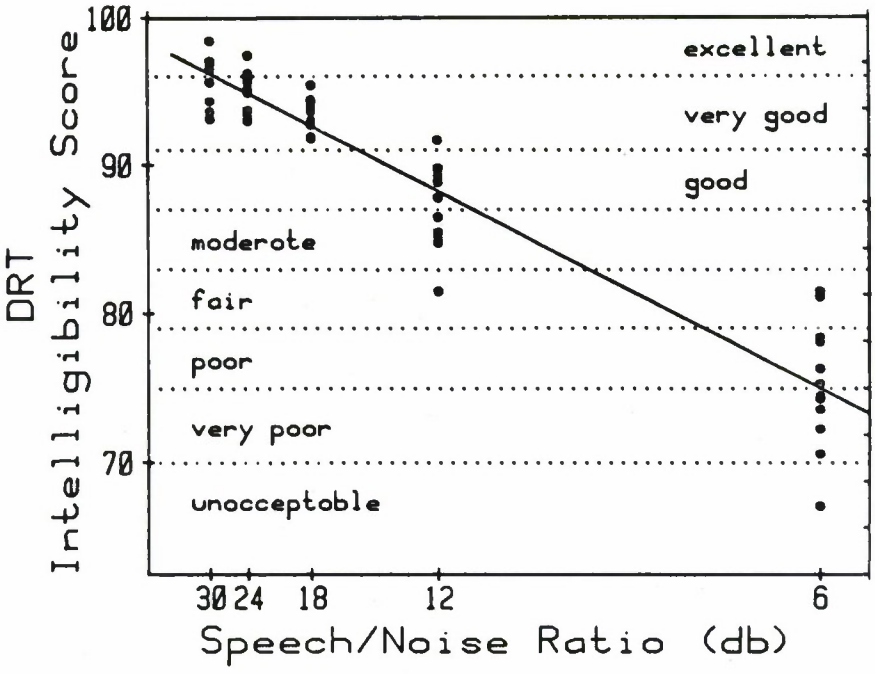

Figure 3. Intelligibility Scores and Regression Model for Speech in Broad-band Noise, in Relation to the Category Scale for Diagnostic Rhyme Test Intelligibility Scores

Individual talkers have been consistently found in hundreds of tests to exhibit significant differences in intelligibility scores (any intelligibility score based on a single speaker should be viewed with suspicion). Significant differences were found in the scores for these three speakers, though the background noise added to the dispersion of the scores and made the speaker differences less conspicuous. The relative ranking of the speaker's scores tended to be maintained each S/N ratio; consequently an alternative regression model with separate regression lines calculated for each speaker and having a common slope resulted in an increase of r^2 to 0.94 and a reduction of the mean square residual. The alternative regression model is illustrated in relation to the scatter diagram of scores in Figure 4.

A "fringe benefit" of intelligibility testing with the Diagnostic Rhyme Test is that it provides separate, independent scores for various phonetic features that contribute to intelligibility[7,8] and permits an evaluation of the effects of noise on

7. Voiers, W.D. (1977) Diagnostic evaluation of speech intelligibility, in Speech Intelligibility and Speaker Recognition, M. Hawley, Ed., Dowden Hutchinson & Ross, Stroudsburg, PA, pp. 374-387.

8. Voiers, W.D. (1983) Evaluating processed speech using the Diagnostic Rhyme Test, Speech Technology, 1(No. 3):30-39.

various components of intelligibility. Those findings are summarized in Figure 5. Nasality was the least impaired by noise interference, followed in order of increasing susceptibility to noise by voicing, compactness, and sibilation; graveness and sustention (the feature that distinguishes between sustained and abrupt consonants) were the features most vulnerable to noise interference. The detailed effects of noise interference on various combinations of feature states, for example, the present and absent states, and the voiced and unvoiced states of the various features, are detailed in Appendix A.

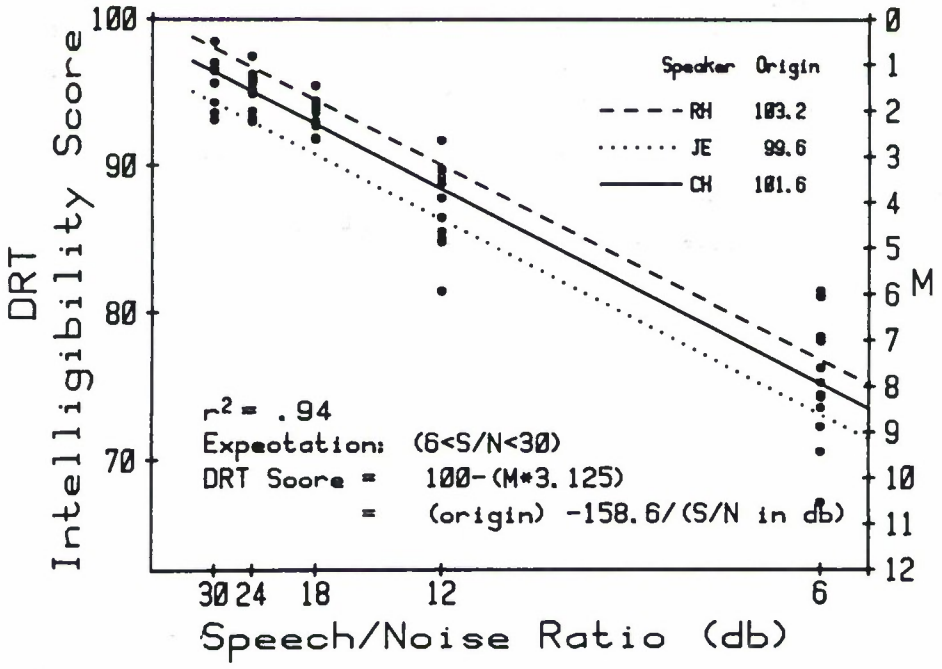

Figure 4. Intelligibility Scores vs S/N Ratio, With Multiple Regression Lines Calculated for Individual Male Speakers

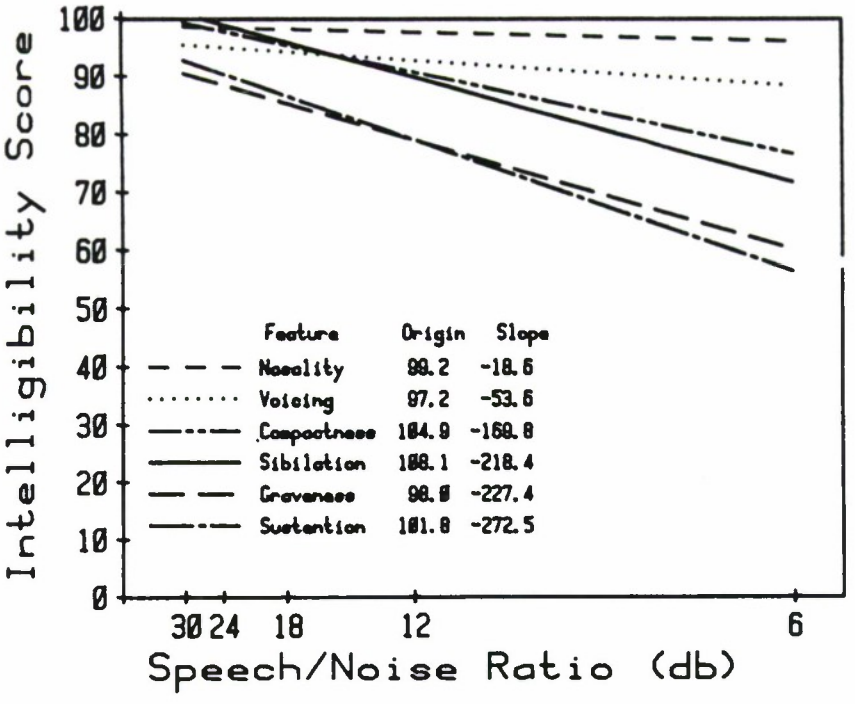

Figure 5. Regression Lines vs S/N Ratio for the Individual Phonetic Features That Contribute to Intelligibility

5. COMPARISONS WITH THE ARTICULATION INDEX

The articulation index provides a means of predicting speech intelligibility of different types of speech materials (nonsense syllables, phonetically balanced word lists, sentences) in relation to the speech signal level and the interfering noise level and their energy spectra, in combination with different listening conditions.[9] Those relationships have been summarized in an American National Standard[10] that provided the basis for the summary shown in Figure 6. The curve labeled "Rhyme Tests" in the figure is based on earlier versions of rhyme tests that, unlike the Diagnostic Rhyme Test, did not provide an adjustment of intelligibility scores for chance effects

9. Beranek, L. L. (1947) The design of speech communications systems IRE Proc., 35(No. 9):880-890.

10. ANSI S3.5-1969 (1969) American National Standard Methods for the Calculation of the Articulation Index, American National Standards Institute, N. Y.

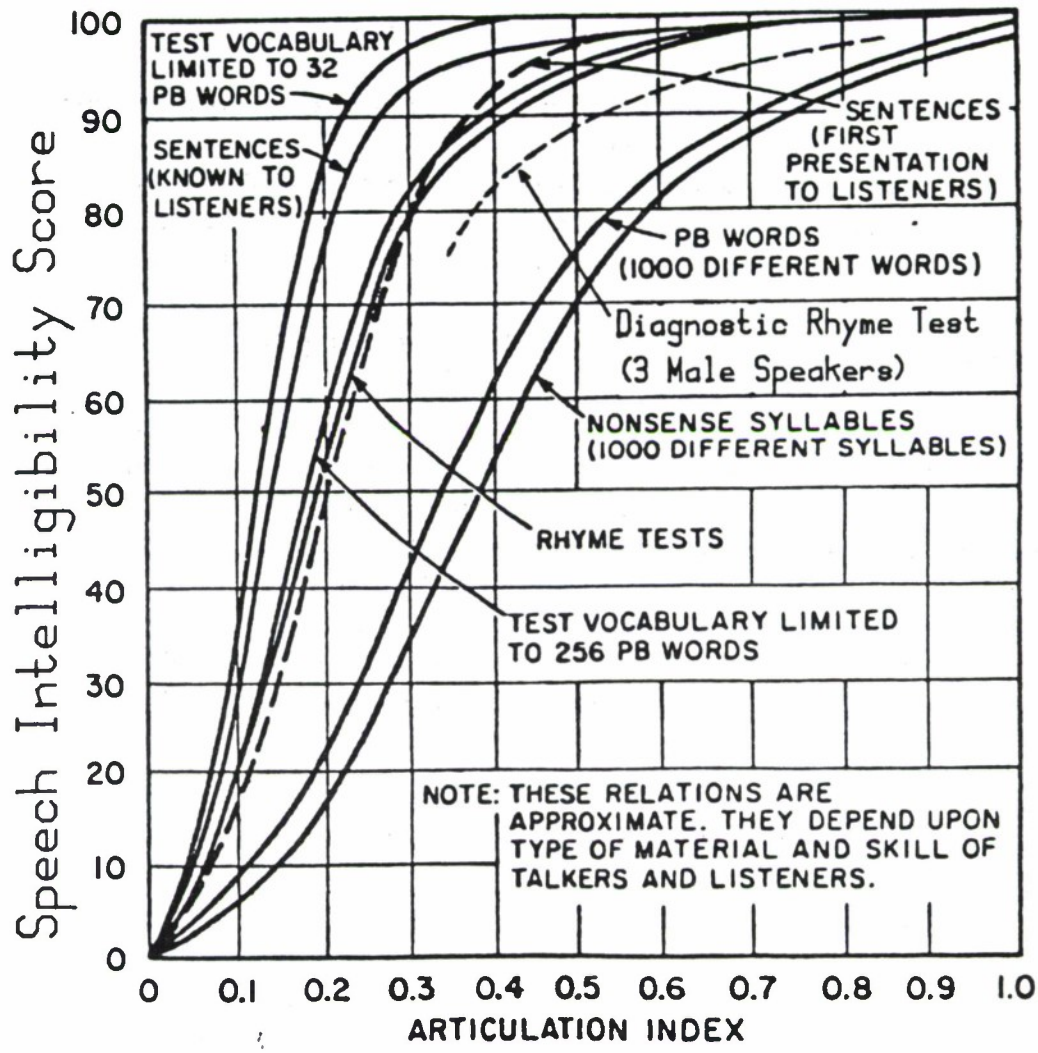

Figure 6. Intelligibility Contours for Different Speech Materials Plotted vs the Articulation Index, With Contour for Diagnostic Rhyme Test Scores From These Studies. This figure derives from ANSI S3.5-1969, American National Standard Methods for the Calculation of the Articulation Index. This version is non-standard, in that the Diagnostic Rhyme Test curve has been added. Also the ordinate scale which is usually labeled "Percent of syllables, words, or sentences understood correctly" is relabeled "Speech Intelligibility Score", as DRT scores are not "percent correct" but include a correction for a priori probabilities

This study made it possible to estimate values of the articulation index at each of the S/N ratios of these tests and construct a new curve that has been added to the figure, estimating the variation in Diagnostic Rhyme Test scores with values of the articulation index over the range studied here.

The ordinate scale in Figure 6 has customarily been labeled "Percent of syllables, words, or sentences understood correctly." However, Diagnostic Rhyme Test scores are corrected for a priori probabilities with the calculation

$$\text{DRT score} = \frac{(\text{Number of items right} - \text{Number of items wrong})}{\text{Total number of items}} \times 100 \ .$$

Accordingly, the ordinate scale in Figure 6 was re-labeled "Speech Intelligibility Score." The dashed contour labeled "Diagnostic Rhyme Test" (three male speakers) represents the relationship calculated in this study. If these DRT scores were modified to remove the correction for chance and thus express "Percent correct" a curve would be obtained that approximates the older curve labeled "Rhyme Tests" for the range investigated here. However, the asymptote of the curve would not approach 100 for "perfect conditions" (A.I. = 1.0) as there are typically about 2 percent listener errors for Diagnostic Rhyme Tests conducted with high fidelity speech signals.

6. SPEECH PERFORMANCE WITH "OPERATIONAL MESSAGES"

The relationships between speech intelligibility and the articulation index summarized in Figure 6 permit estimation of speech performance with stereotyped voice messages well-known to listeners (typical "operational messages" that have been advocated for use in speech test and evaluation), shown in Table 1.

Extrapolation of the curve for DRT scores vs the AI gives an estimate of an articulation index of about 0.30 for a DRT score of 70, the score that has been postulated as representing a threshold score representing the boundary between "unacceptable" and "very poor" intelligibility performance. While the ANSI relationships shown in Figure 6 suggest that well-known stereotyped voice messages might be received with better than 90 percent accuracy under such conditions, the relationships also suggest that were any emergency to occur in which communicators were required to depart from their usual stereotyped communications and need to use unfamilar words and phrases, the intelligibility performance of that channel would present serious difficulties. Figure 6 also tends to explain why some speech tests using "operational messages" resulted in subjects producing judgements that a voice processor had acceptable performance, even though that processor had scored below 70 in formal Diagnostic Rhyme Tests.

Table 1. Estimate of Speech Performance With "Operational Messages" Based on the Articulation Index

S/N Ratio	Avg. DRT Score	Estimated A. I.	Est. of Avg. percent correct: Sentences known to Listeners ('Operational Messages')
30 db	96	.85	99%
24 db	95	.76	99%
18 db	93	.64	98%
12 db	87	.50	97%
6 db	75	.34	94%
(by extrapolation:)			
	(70)	(.30)	(92%)

7. VOICE QUALITY AND ACCEPTABILITY TESTS

As the voice quality and acceptability tests were not replicated there were far fewer listener scores than for intelligibility. Separate scores for signal quality and for background quality are obtained with the Diagnostic Acceptability Measure (DAM) test.[3] A weighted combination of background and signal quality scores produces scores for overall quality, called the Composite Acceptability Estimate (DAM/CAE). With additive background and linear processing it might be anticipated that background quality scores would vary with the S/N ratio and signal quality scores would remain relatively constant. This was not the case. While background quality scores varied more widely, there was also significant variation in the judgements of signal quality at the various background noise levels. Figure 7 shows the scatter diagram of signal quality scores (DAM/CSA) in relation to S/N ratio with the regression line and 95 percent confidence limits for the ensemble of scores, modelled in relation to the reciprocal of S/N ratio.

Scores for background quality, representing the Composite Background Acceptability (DAM/CBA) are shown in relation to S/N ratio in Figure 8. A 2nd order regression model based on the reciprocal of S/N ratio resulted in a value of $r^2 = 0.98$ and a standard error of estimate of 1.38, with a range of approximately 25 compared with a range of approximately 10 points for signal quality.

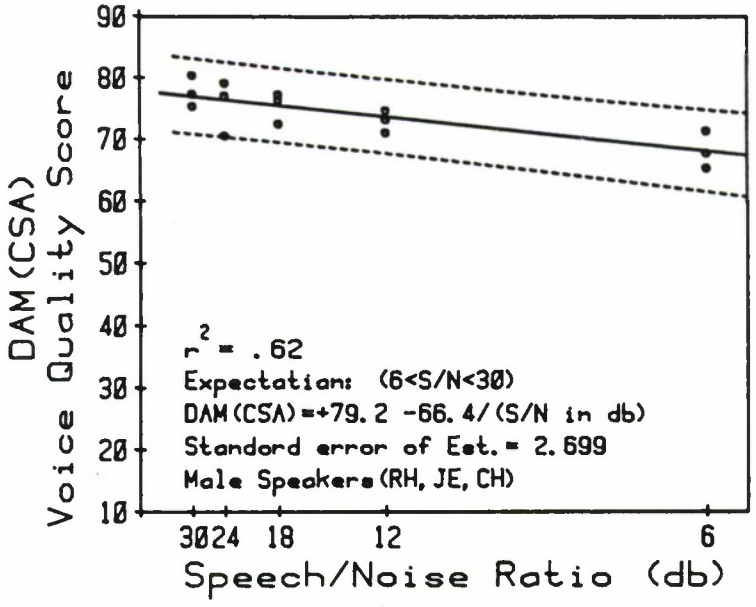

Figure 7. Scatter Diagram of Scores for Signal Quality With Regression Line and 95 Percent Confidence Limits for the Data Points, Based on Reciprocal of S/N Ratio

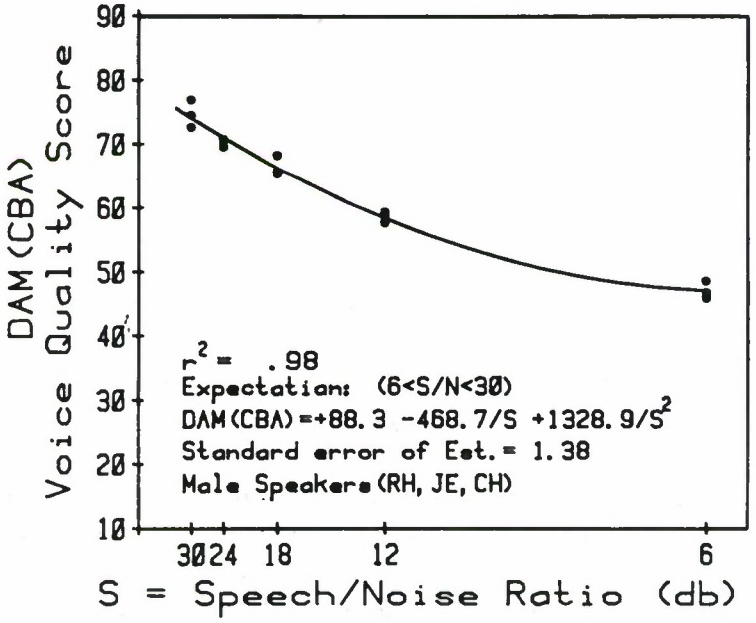

Figure 8. Scatter Diagram of Scores for Background Quality vs S/N Ratio With 2nd Order Regression Model Based on the Reciprocal of S/N Ratio

Scores for overall quality are not the average of the signal and background quality scores, but a function of the product of the two. A scatter diagram of those scores representing the Composite Acceptability Estimate (DAM/CAE) is presented in Figure 9 with a 2nd order regression curve based on the reciprocal of S/N ratio. This regression model resulted in a value of $r^2 = 0.96$ and a standard error of estimate of 1.94.

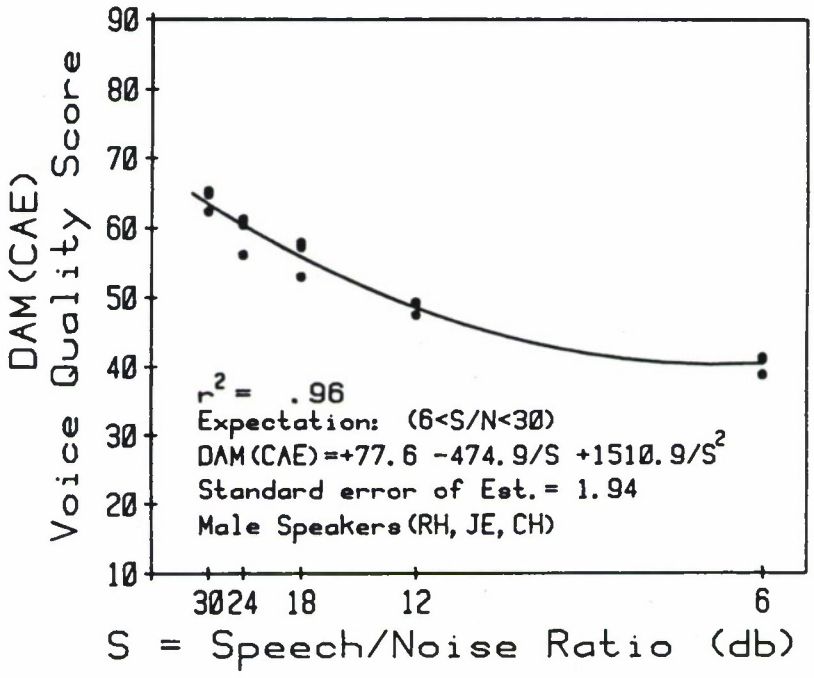

Figure 9. Scatter Diagram of Scores for Overall Quality vs S/N Ratio, With a 2nd Order Regression Model Based on the Reciprocal of S/N Ratio

The same caveat regarding extrapolation to estimate scores outside the measured range of S/N ratios (6 dB to 30 dB) expressed for speech intelligibility data applies here, and even more strongly in the case of the 2nd order regression models. Again, however, the regression model serves a useful purpose in conjunction with a scale of categories that has been established to assist in the interpretation of DAM voice quality ratings. The category scale, shown in Figure 10, utilizes the same eight labels for categories as used for intelligibility scores, ranging from "excellent" to "unacceptable". The category scale refers to the scores for overall voice quality (DAM/CAE); no separate category scales for signal and background quality have been attempted.

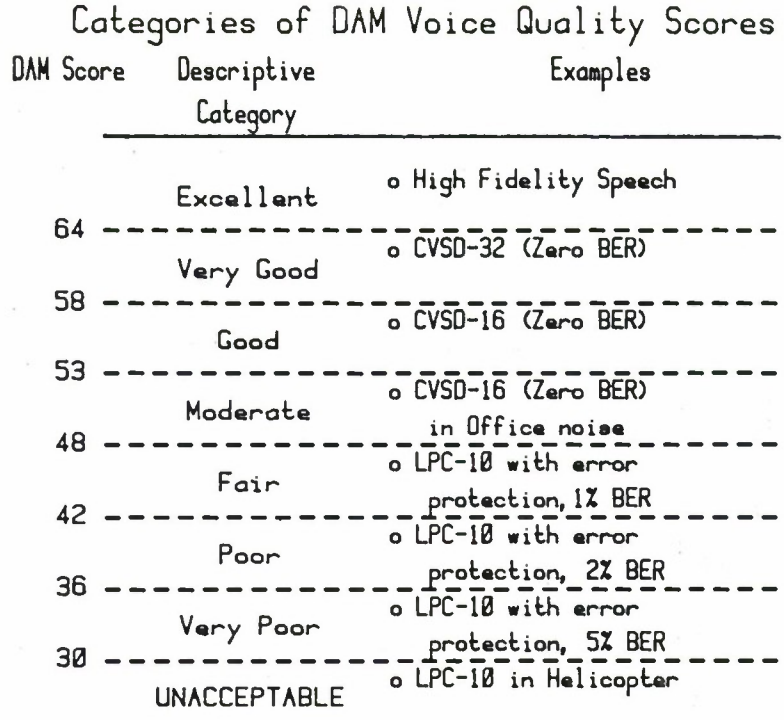

Figure 10. Category Scale for Diagnostic Acceptability Measure Voice Quality Scores With Examples of Voice Processor Categories

As with the category scale for intelligibility scores, these labels do not represent judgements of listeners but judgements of a committee of experts that has been involved with extensive tests of voice processors and has had opportunities to obtain informal judgements of voice processor quality from users and correlate those opinions with results of formal DAM tests. This category scale should be considered tentative and subject to revision as further knowledge is gained (as is the case with the category scale for intelligibility scores).

Combining the category scale for voice quality scores with the data obtained in these studies produce the result shown in Figure 11.

The 30 dB S/N ratio resulted in voice quality ratings clustered around the boundary between the "excellent" and "very good" categories, a result that by coincidence was similar to that obtained with speech intelligibility scores. The 24 dB S/N ratio resulted in voice quality ratings in the "very good" and "good" categories, while the 18 dB S/N ratio resulted in scores bracketing the "good" category. At 12 dB S/N ratio the voice quality scores clustered around the boundary between "moderate" and "fair". The 6 dB S/N ratio produced scores in the "poor" category.

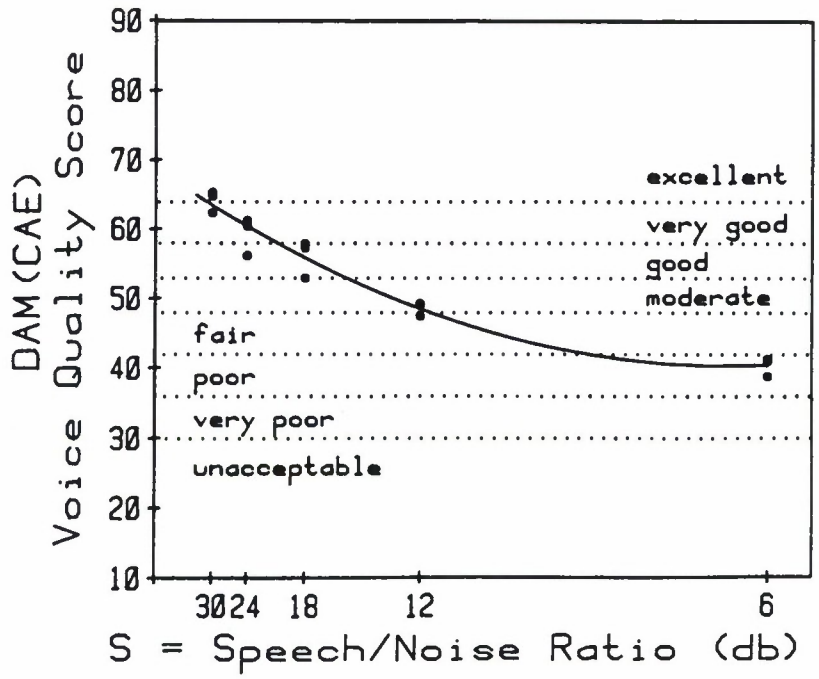

Figure 11. Diagnostic Acceptability Measure Scores for Overall Quality vs S/N Ratio and Regression Model, in Relation to the Category Scale for Diagnostic Acceptability Measure Voice Quality Scores

8. MEAN OPINION SCORES

Mean opinion scores result from tests in which listeners make direct judgements of telephony channels. There are differing versions of the test procedure. In one version, subjects conduct conversations over the telephony channel under test and then make their judgements of the channel. Other versions involve only listening to speech samples and then rating the speech sample. In this instance the ratings were obtained by the latter method, using a five-point scale representing "excellent", "good", "fair", "poor", and "bad", with results shown in Figure 12 together with a regression model.

The speech samples on which listeners made their judgements were the same sentence recordings used for the Diagnostic Acceptability Measure voice quality tests. The regression model best fitting these points was based on the S/N ratios rather than the reciprocals of those values, and resulted in a value of $r^2 = 0.93$ and a standard error of estimate of 0.17. The ratings and the regression curve are shown in relation to the category scale for the mean opinion scores in Figure 13.

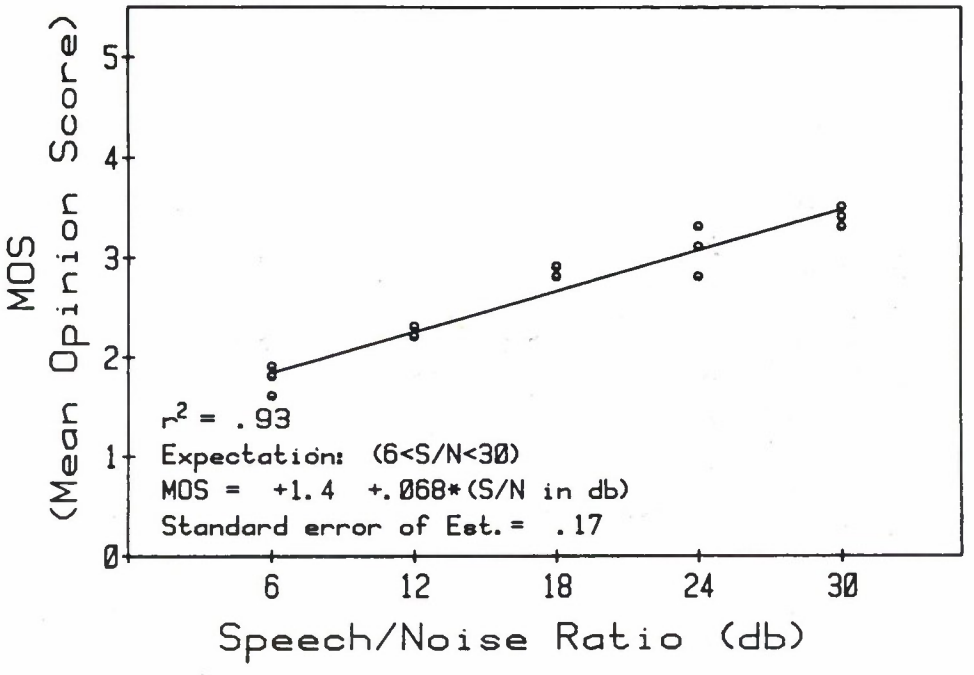

Figure 12. Scatter Diagram of Mean Opinion Scores vs S/N Ratio With Linear Regression Model

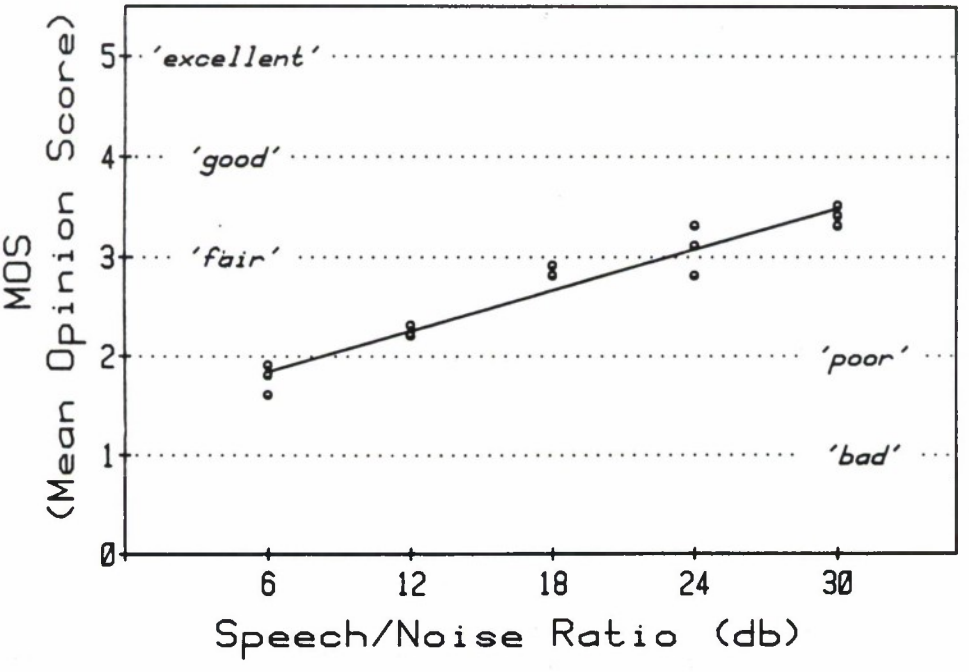

Figure 13. Mean Opinion Scores and Linear Regression Model in Relation to the Categories Used by Listeners

9. COMPARISON OF CATEGORY SCALES

In Figure 14, the category scales for intelligibility and quality ratings are compared, based on their common relationship to S/N ratio; the values calculated for the articulation index at the five S/N ratios tested are also shown.

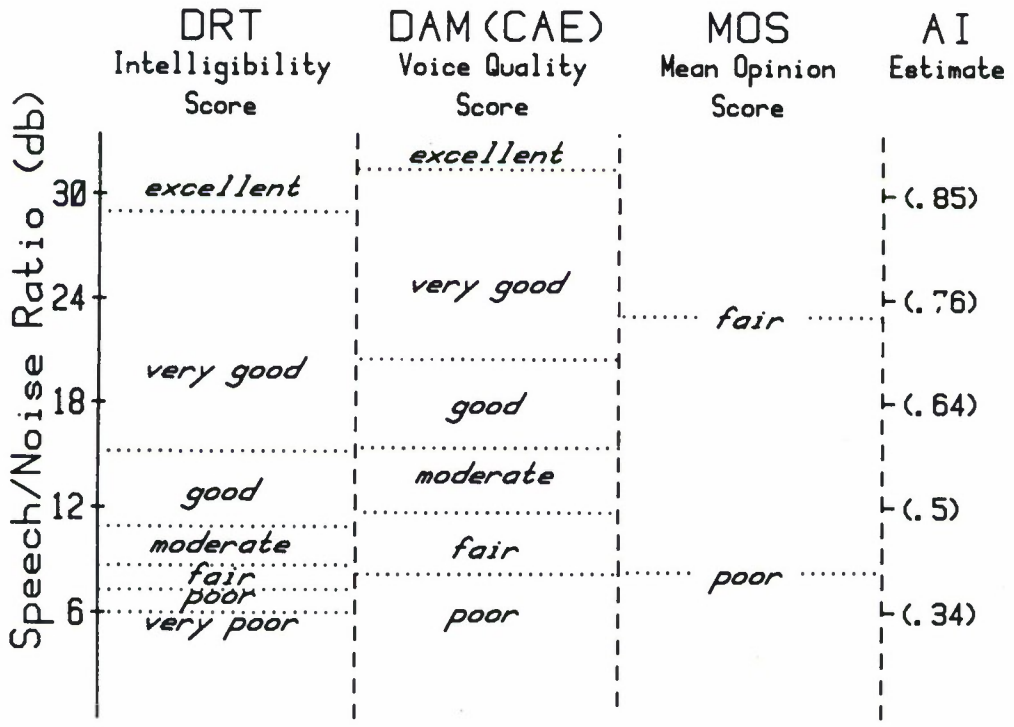

Figure 14. Comparison of the Category Scales for the Intelligibility Scores, Voice Quality Scores, and Mean Opinion Scores in Relation to the Scale for S/N Ratio and the Estimates of Values of the Articulation Index

Diagnostic Rhyme Test intelligibility and Diagnostic Acceptability Measure voice quality scales are in fair agreement at the top and bottom of the range but show little agreement in the middle of the range. The mean opinion score ratings gave fair agreement only at the bottom of this scale.

10. DISCUSSION

The discrepancies between the category scales for Diagnostic Rhyme Test intelligibility scores, Diagnostic Acceptability Measurement voice quality scores, and Mean Opinion Scores emphasize the different origins of these scales. While the categories representing the numerical ratings in obtaining Mean Opinion Scores represent direct judgements by listener crews, the other two category scales were created by a committee with members long experienced in test and evaluation of digital voice communications processors and the interpretation of Diagnostic Rhyme Test Scores and Diagnostic Acceptability Measurement scores. Repeatedly faced with the problem of interpreting to others the significance of scores obtained for voice processors under different test conditions, it was decided to construct rating scales based on descriptive labels, that might be used by anyone wishing to estimate the significance of a particular Diagnostic Rhyme Test intelligibility score or Diagnostic Acceptability Measurement voice quality rating. The category scales for these scores were constructed by the committee after many discussions of this issue and extensive reviews of performance data covering a wide variety of processors and test conditions.

It is therefore not surprising, considering the ad hoc nature of the Diagnostic Rhyme Test and Diagnostic Acceptability Measurement category scales, that the categorizations do not agree well in their relationship to the effects of broad-band noise on speech. These findings may provide the basis and incentive for further studies of the discrepancies between the category scales and the issue of whether the scales might or should be brought into closer agreement.

11. CONCLUSIONS

Tests of speech with additive broad-band noise resulted in the following findings:

- Estimates were made of the relationship between Diagnostic Rhyme Test intelligibility scores, and values of the articulation index.
- A comparison of category scales for Diagnostic Rhyme Test intelligibility scores, Diagnostic Acceptability Measurement voice quality ratings, and Mean Opinion Scores was established. Values of the articulation index in relation to those category scales were established.

- From the relation between Diagnostic Rhyme Test scores and the articulation index it was possible to make estimates of the percent correct of stereotyped messages known to listeners, that is, "operational messages", in relation to Diagnostic Rhyme Test scores.
- Test results highlighted the importance of conducting speech test and evaluation with multiple speakers, and of replicating tests whenever practicable.
- The study confirmed the utility of the new algorithm developed in the Rome Air Development Center speech test and evaluation facility for measuring and calibrating speech-to-noise energy ratios.

References

1. Sims, J.T. (1985) A speech-to-noise ratio measurement algorithm, J. Acoust. Soc. Am., 78(No. 5):1671-1674.
2. Brady, P.T. (1968) Equivalent peak level: a threshold-independent speech-level measure, J. Acoust. Soc. Am., 44:695-699.
3. Voiers, W.D. (1977) Diagnostic acceptability measure for speech communications systems, IEEE Proc. ICASSP 77CH1197-3 ASSP, pp. 204-207.
4. CCITT (1981) Telephone Transmission Quality: Recommendations of the P Series, Yellow Book Vol. V, ITU, Geneva.
5. Smith, C.P. (1983) Narrowband (LPC-10) Vocoder Performance Under Combined Effects of Random Bit Errors and Jet Aircraft Cabin Noise, RADC-TR-83-293, AD A141333, Rome Air Development Center, Griffiss AFB, N.Y.
6. Smith, C.P. (1983) Relating the performance of speech processor to the bit error rate, Speech Technology 2(No. 1):41-53.
7. Voiers, W.D. (1977) Diagnostic evaluation of speech intelligibility, in Speech Intelligibility and Speaker Recognition, M. Hawley, Ed., Dowden Hutchinson & Ross, Stroudsburg, PA, pp. 374-387.
8. Voiers, W.D. (1983) Evaluating processed speech using the Diagnostic Rhyme Test, Speech Technology, 1(No. 3):30-39.
9. Beranek, L.L. (1947) The design of speech communications systems, IRE Proc., 35(No. 9):880-890.
10. ANSI S3.5-1969 (1969) American National Standard Methods for the Calculation of the Articulation Index, American National Standards Institute, N.Y.

Appendix A

Details of Variation in Diagnostic Rhyme Test Feature Scores With Speech-to-Noise Energy Ratios

Figures A1 through A15 that follow, present the results of analyzing separately the effects of broad-band noise on each of the intelligibility feature states.

Nasality was little affected by noise over the range tested here; this was true not only for the overall scores for this feature, but also for the contrasts between nasality-present and nasality-absent, and for the grave and acute states of this intelligibility feature.

The remaining feature scores tended to show varying degrees of susceptibility to noise interference. The voiced state of sibilation was degraded by noise to a greater degree than the unvoiced state; however the opposite was true for the features graveness and compactness.

The present state of sibilition and graveness were more susceptible to noise than the absent state; however the opposite was true of the features voicing and compactness. The feature scores that exhibited significant differences among speakers included voicing (frictional) and voicing (total); sustention (voiced), sustention (unvoiced) and sustention (total).

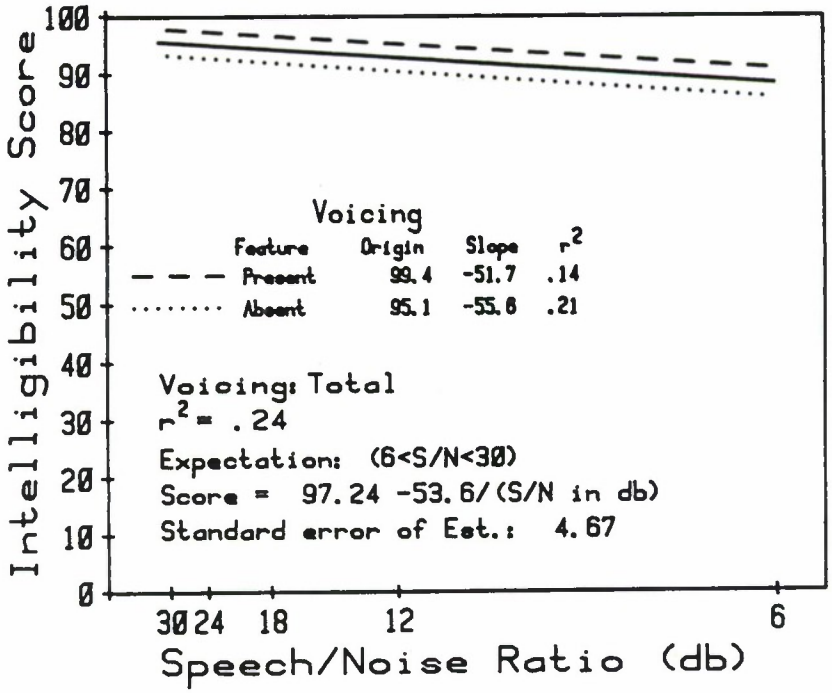

Figure A1. Regression Models for the Scores for Voicing-present, Voicing-absent, and Voicing (total) vs the Reciprocal of S/N Ratio

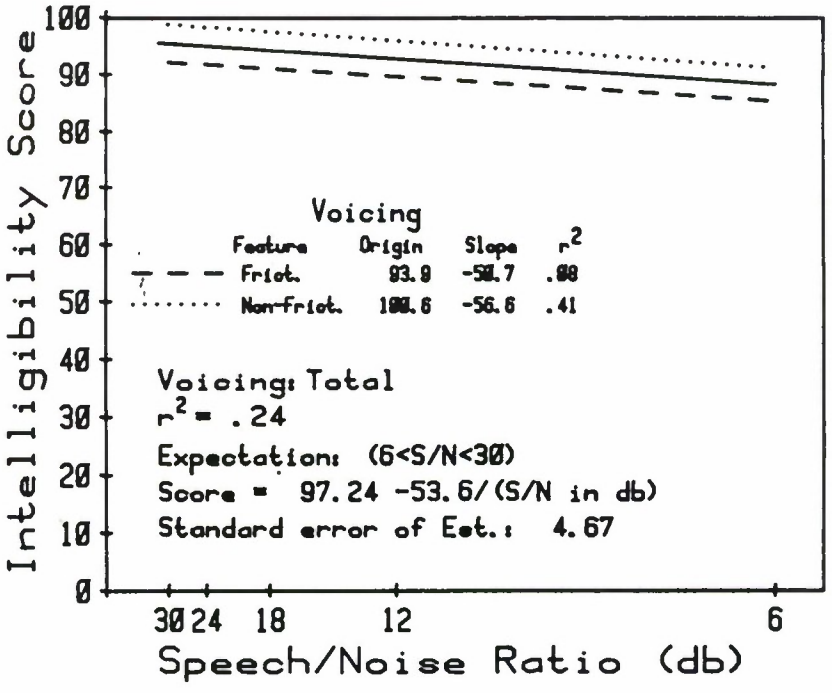

Figure A2. Regression Models for Voicing-frictional, Voicing-non-frictional, and Voicing (total) vs the Reciprocal of S/N Ratio

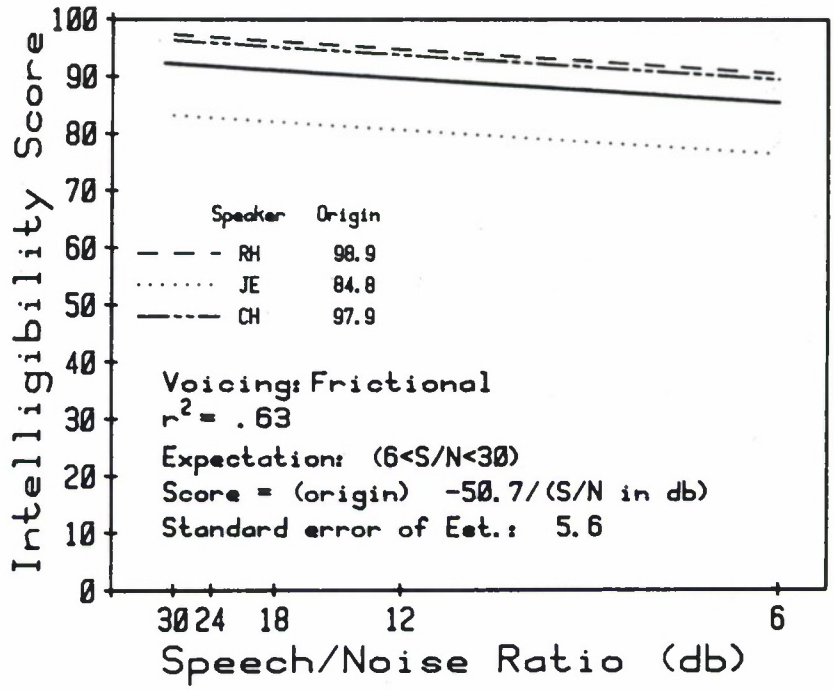

Figure A3. Regression Models for Voicing-frictional vs the Reciprocal of S/N Ratio, Indicating the Differences Among the Three Male Speakers

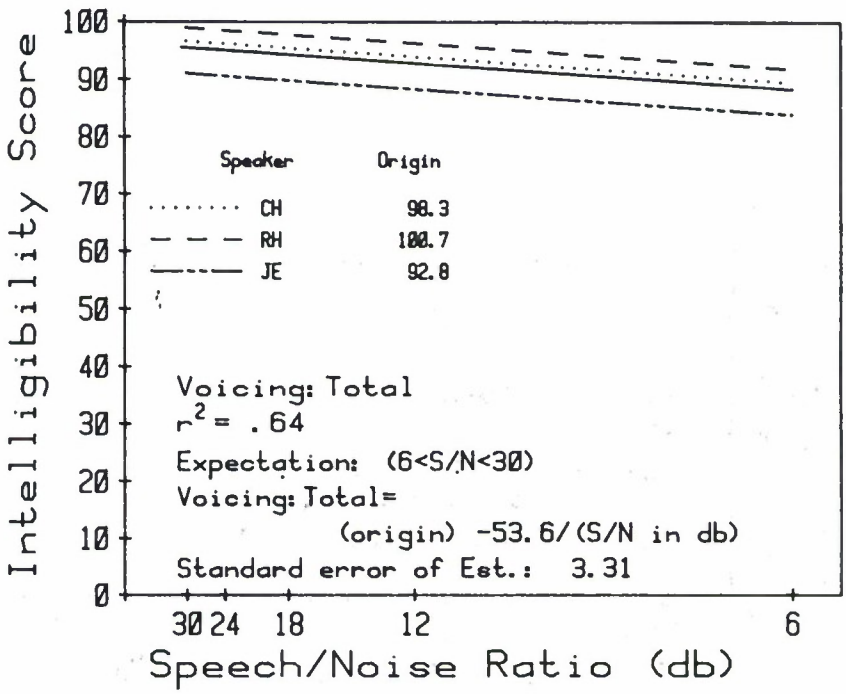

Figure A4. Regression Models for Voicing (total) vs the Reciprocal for S/N Ratio, Indicating Differences Among the Three Male Speakers

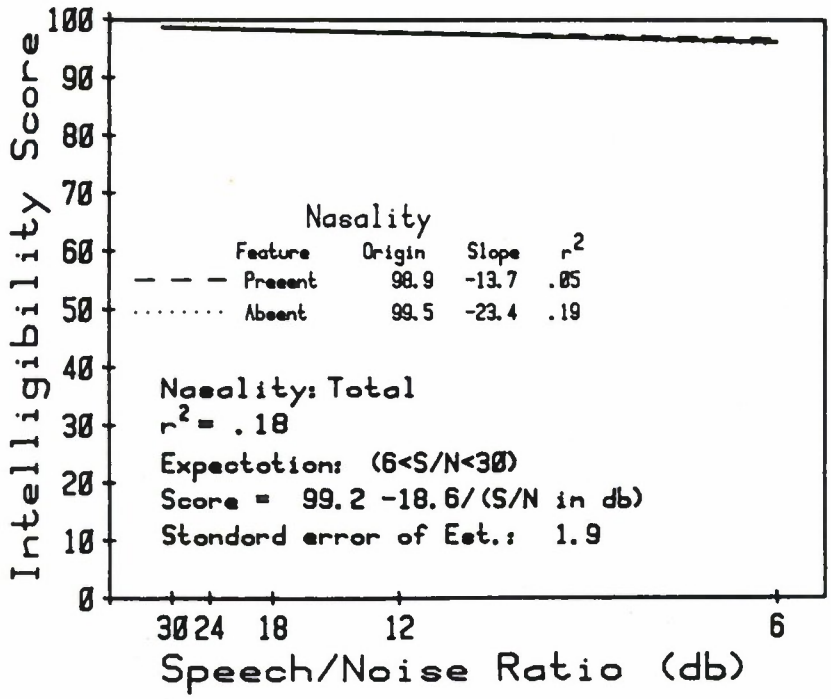

Figure A5. Regression Models for Nasality Scores vs S/N Ratio, Indicating Virtually No Differences Between the Total Score, and the Present and Absent State

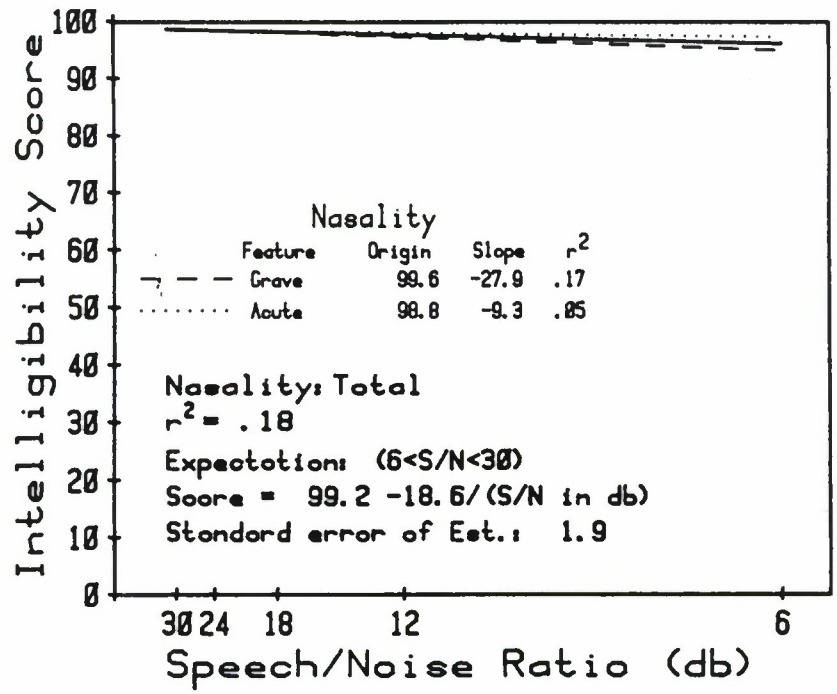

Figure A6. Regression Models for Nasality Scores vs S/N Ratio, Indicating Only Slight Differences Between Total Scores and the Grave and Acute States

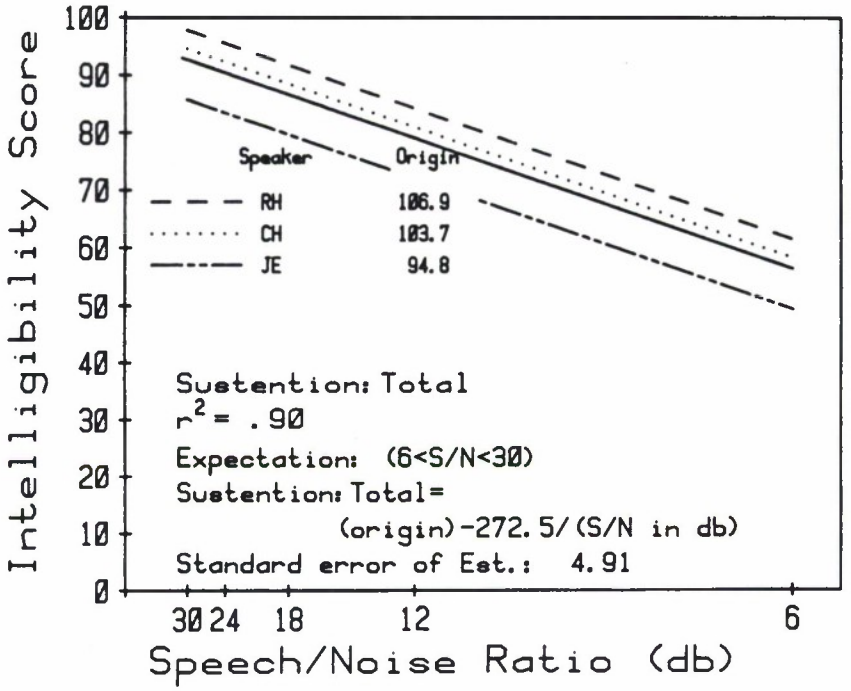

Figure A7. Regression Models for Sustention (total) vs Reciprocal of S/N Ratio Showing Differences Among the Three Male Speakers

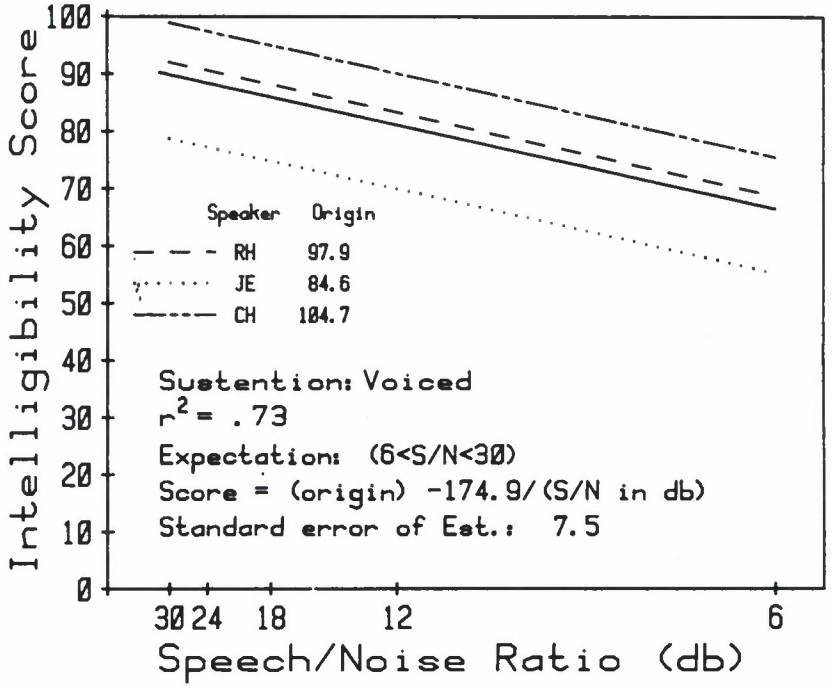

Figure A8. Regression Models for Sustention-voiced vs the Reciprocal of S/N Ratio Showing Differences Among the Three Male Speakers

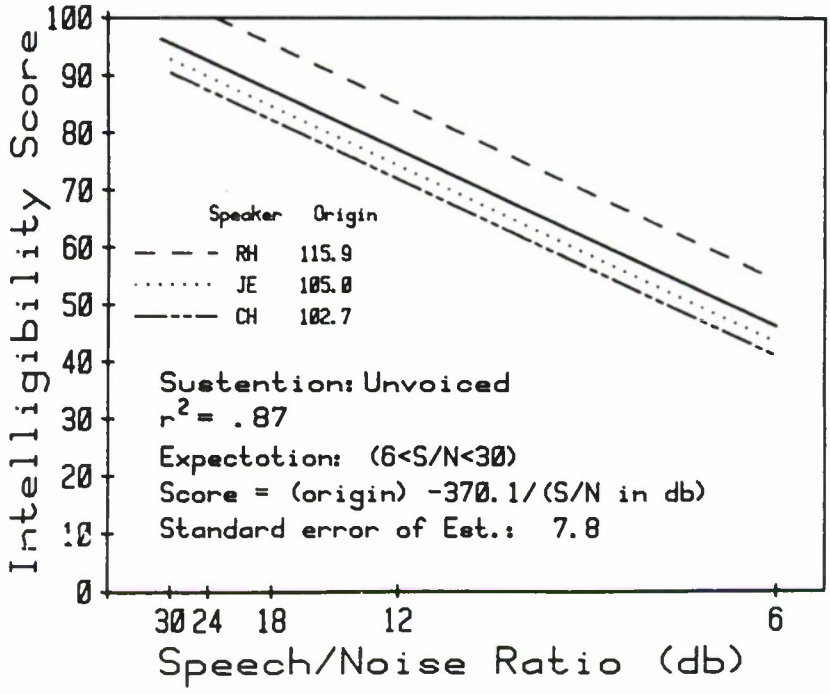

Figure A9. Regression Models for Sustention-unvoiced vs Reciprocal of S/N Ratio Showing Differences Among the Three Male Speakers

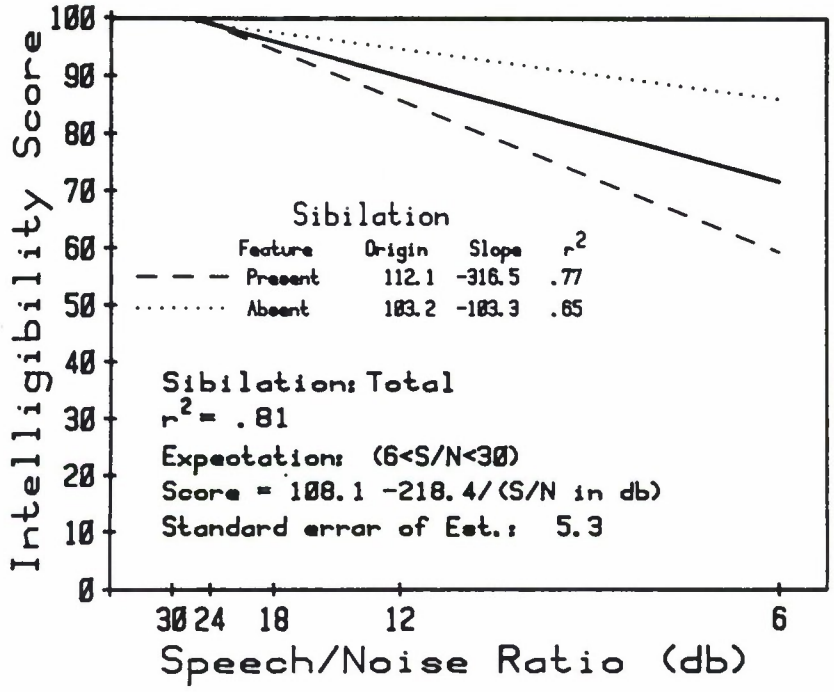

Figure A10. Regression Models for Sibilation vs the Reciprocal of S/N Ratio, Showing the Differences Between the Present and Absent State of Sibilation

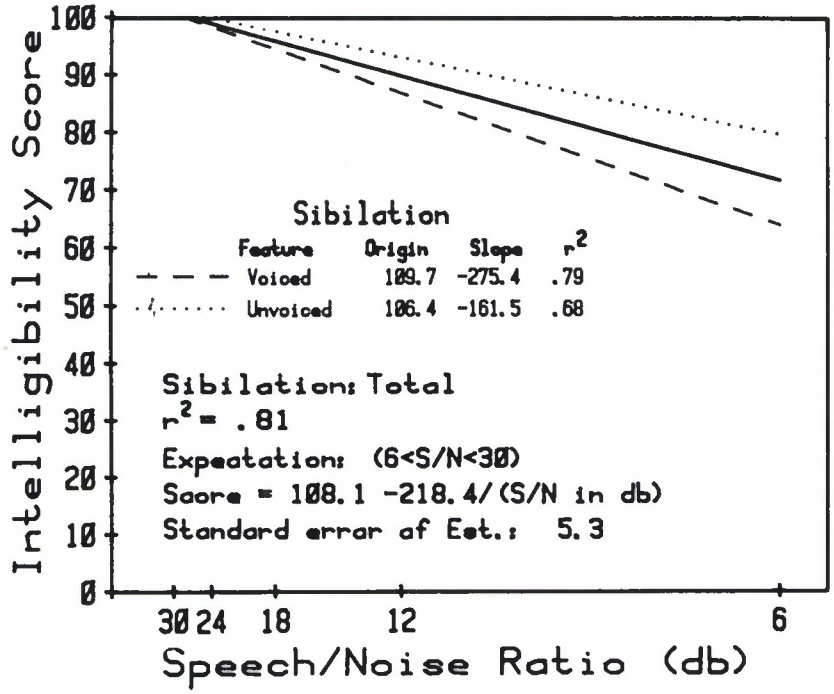

Figure A11. Regression Models for Sibilation vs the Reciprocal of S/N Ratio, Showing the Differences Between the Voiced and Unvoiced State of Sibilation

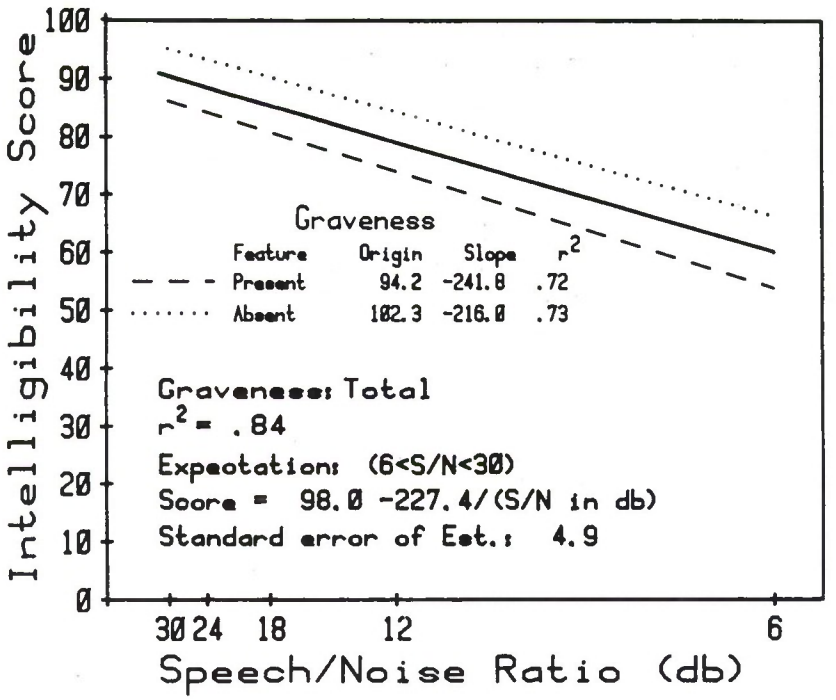

Figure 12. Regression Models for Graveness vs the Reciprocal of S/N Ratio, Showing the Differences Between the Present and Absent State of the Graveness Feature

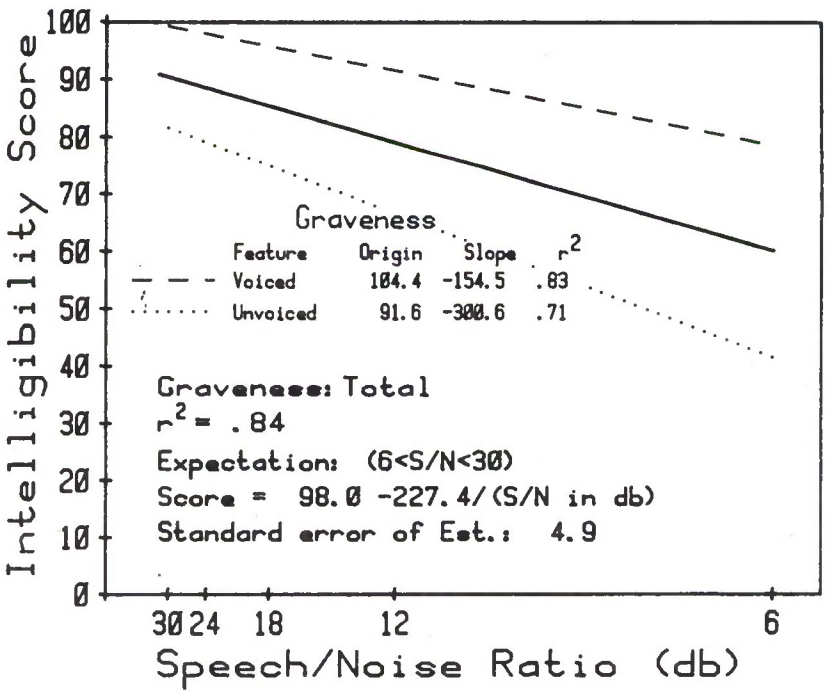

Figure A13. Regression Models for Graveness vs the Reciprocal of S/N Ratio, Showing the Differences Between the Voiced and Unvoiced State of the Graveness Feature

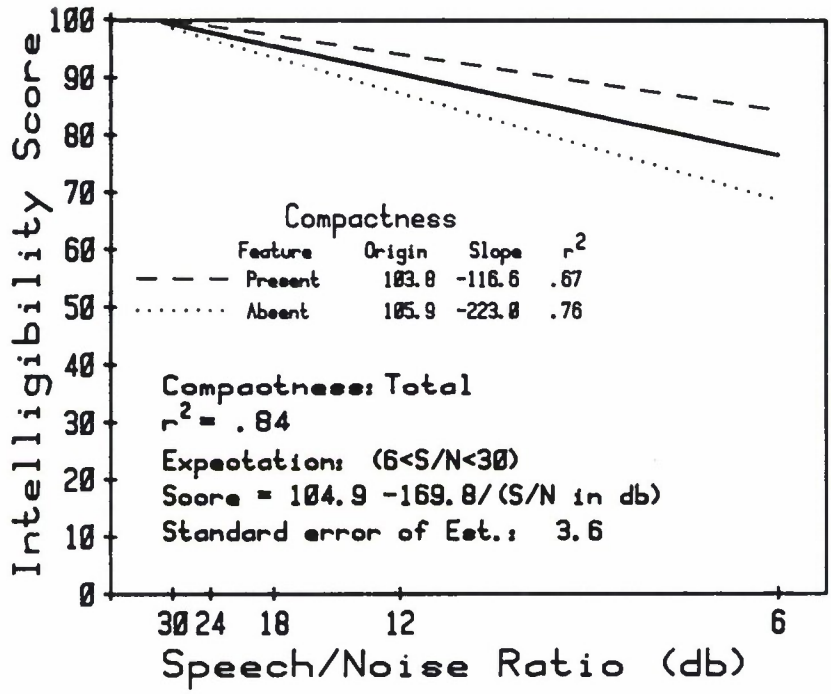

Figure 14. Regression Models for Compactness vs the Reciprocal of S/N Ratio, Showing the Differences Between the Present and Absent State of the Compactness Feature

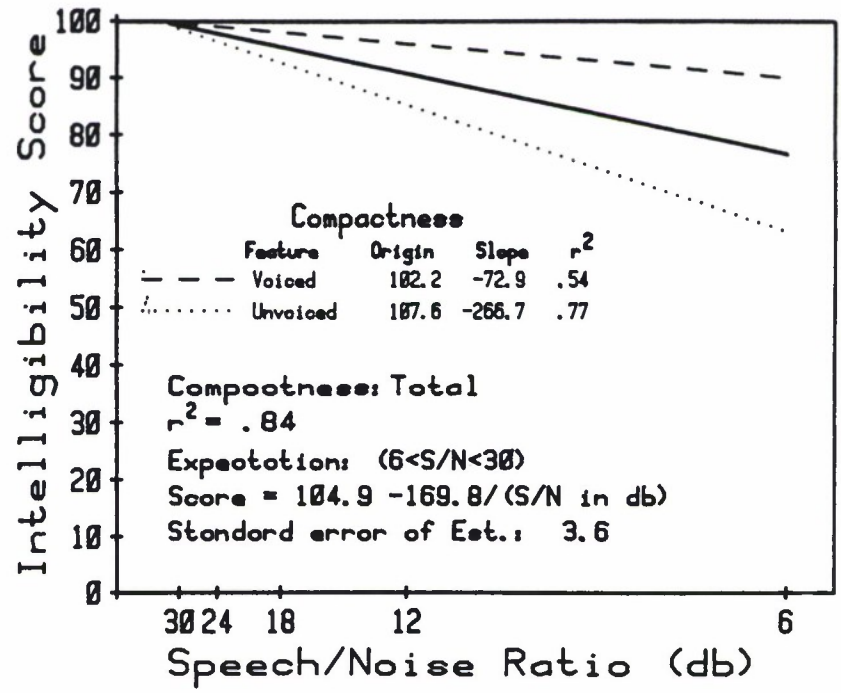

Figure A15. Regression Models for Compactness vs the Reciprocal of S/N Ratio, Showing the Differences Between the Voiced and Unvoiced State of the Compactness Feature